COMPARISON OF WINDOWING SCHEMES FOR SPEECH CODING

Johannes Fischer[*] *and Tom Bäckström*[*†]

[*]International Audio Laboratories Erlangen, Friedrich-Alexander-University Erlangen-Nürnberg (FAU)
[†]Fraunhofer IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany

johannes.fischer@audiolabs-erlangen.de

ABSTRACT

The majority of speech coding algorithms are based on the code excited linear prediction (CELP) paradigm, modelling the speech signal by linear prediction. This coding approach offers the advantage of a very short algorithmic delay, due to the windowing scheme based on rectangular windowing of the residual of the linear predictor. Although widely used, the performance and structural choices of this windowing scheme have not been extensively documented. In this paper we introduce three alternative windowing schemes, as alternatives to the one already used in CELP codecs. These windowing schemes differ in their handling of transitions between frames. Our subject evaluation shows that omitting the error feedback loop yields an increase in perceptual quality at scenarios with high quantization noise. In addition, objective measures show that while error feedback improves the accuracy slightly at high bitrates, at low bitrates it causes a degradation in quality, resulting in a lower SNR.

Index Terms— speech coding, windowing, source modelling, linear prediction

1. INTRODUCTION

Windowing of the input signal is necessary in most speech processing applications, including coding of the speech signal in the time domain using CELP type codecs or processing of the speech signal in the frequency domain. This process of segmentation usually constitutes the first step of the algorithms. The most common approach is the overlap-add concept, the performance and properties of which are well-known and understood [1, 2].

Speech coders model the signal with a linear predictor, such that the residual can be windowed with a rectangular window as described in preceding work [3, 4]. Although, this filter-windowing is applied in all mainstream speech codecs as AMR-WB [5], G.718 [6], MPEG USAC [7] and EVS [8], this combination of filtering and windowing has not been documented in detail. Whereas other components of CELP-coders have been optimized and documented thoroughly, publications pay little attention to the windowing scheme that is inherently applied.

This windowing scheme is described amongst others in [3], and a minimum mean square error (MMSE) solution for quantization is given. In contrast to the available work, that evaluates the performance of the whole speech coder, the goal of this paper is to evaluate the windowing performance in an isolated fashion. Other publications [9, 10] improved the estimation of linear predictive coding (LPC) coefficients by applying similar kind of windows. However, the presented windowing scheme is only applicable as an analysis window. In contrast, the windowing schemes described in this paper are designed for general speech processing and coding purposes.

A first review of windowing schemes based on filters, which we call filter-windowing, and overlap-add windowing was presented in [11]. In this paper we present and evaluate three modifications of the windowing scheme in [11] and compare them to the commonly used scheme of code excited linear prediction (CELP) codecs. Moreover, an MMSE solution for quantization of the residual is presented for the introduced windowing schemes. The objective of this paper is an isolated evaluation of the performance of the windowing schemes, whereby these were implemented independently from a speech coder. This is necessary not only to focus on the windowing schemes, but also not to give advantage to the original windowing scheme, as the given speech coders are all fine tuned to maximum performance, applying the known windowing scheme. Therefore, quantization was simulated by adding white noise to the residual.

For the objective evaluation, results of the segmental signal-to-noise ratio (SNR) of the output are presented. To evaluate subjective perceptual quality, we conducted a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening test. The experiments show that all windowing schemes offer competitive performance, although the feedback loop, used by many speech coders to optimize SNR degrades performance at high quantization noise levels.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

2. DEFINITIONS AND NOTATION

Speech codecs based on the CELP paradagim express speech signals in terms of a linear predictor and its residual. Therefore, linear prediction of order N is applied to the speech signal $s(t)$, to obtain the residual signal $r(t)$. Given the predition filter $\alpha(t)$, the residual follows by: $r(t) = -s(t) * \alpha(t)$, where $(*)$ denotes the convolution operator. The linear predictive filter is not stationary, but slowly evolving, whereby it can be assumed to be constant over short windows. Therefore, the input speech signal is partitioned into frames of length M, indexed by k. Therefore, $\mathbf{s}_k = [\sigma_{k,0} \; ... \; \sigma_{k,M-1}]$ and $\mathbf{r}_k = [\rho_{k,0} \; ... \; \rho_{k,M-1}]$ denote the kth frame of the input and the residual, respectively. As the prediction filter is changing with each frame k, we denote the time varying filter by $\mathbf{a}_k = [\alpha_{k,0} \; ... \; \alpha_{k,N}]$.

Quantization of the residual is simulated by adding uncorrelated noise to the residual, such that a desired SNR is achieved. Therefore, a noise frame $\mathbf{n}_k = [\nu_{k,0} \; ... \; \nu_{k,M-1}]$ is added to the residual, such that the quantized residual is $\mathring{\mathbf{r}}_k = \mathbf{r}_k + \mathbf{n}_k$. As the convolution matrix $\mathbf{A}_k$

$$\mathbf{A}_k = [\mathbf{U}_k \mathbf{L}_k] = \left[\begin{array}{ccc:cccc} \alpha_{k,N} & \dots & \alpha_{k,1} & 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \alpha_{k,N} & \dots & \alpha_{k,1} & 1 & 0 \\ 0 & \dots & 0 & \alpha_{k,N} & \dots & \alpha_{k,1} & 1 \end{array}\right] \tag{1}$$

consists of the filter coefficients α, multiplication by $\mathbf{A}_k$ corresponds to finite impulse response (FIR) filtering. This convolution matrix can be split into an upper $\mathbf{U}_k$ and a lower $\mathbf{L}_k$ triangular matrix of Toeplitz structure. It is well-known that infinite impulse response (IIR) filtering with a filter $\alpha_{k,h}$ is equivalent to FIR filtering with the impulse response of the filter $\eta_{k,h}$. Thus, we can define a convolution matrix $\mathbf{H}_k$

$$\mathbf{H}_k = \begin{bmatrix} \underline{\mathbf{G}} \\ \mathbf{T}_1 \\ \mathbf{T}_2 \\ \vdots \end{bmatrix} = \left[\begin{array}{cccc} 1 & 0 & \dots & 0 \\ \eta_{k,1} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \eta_{k,N-1} & \dots & \eta_{k,1} & 1 \\ \hdashline \vdots & & \vdots & \vdots \end{array}\right] \tag{2}$$

consisting of the impulse response of the linear predictive filter. Its length is by definition infinite and a multiplication will yield an IIR filter operation. Similarly as matrix $\mathbf{A}_k$, also the matrix $\mathbf{H}_k$ can be divided into a lower triangular matrix $\mathbf{G}$ and matrices $\mathbf{T}$ all being of Toeplitz structure and of size $\mathrm{M} \times \mathrm{M}$.

3. WINDOWING SCHEMES

As ***Conventional IIR***, we will refer in the following to the windowing scheme currently applied in CELP speech coders, as also described in [3]. In this approach, the speech signal is modelled by an IIR filter, whereby the residual signal $\mathbf{r}_k$ is obtained by FIR filtering the input with the linear predictive filter, represented by $\mathbf{A}_k$. Therefore, the residual can be obtained by

$$\begin{aligned} \mathbf{r}_k = \mathbf{A}_k \begin{bmatrix} \mathbf{s}_{k-1} \\ \mathbf{s}_k \end{bmatrix} &= \begin{bmatrix} \mathbf{U}_k & \mathbf{L}_k \end{bmatrix} \begin{bmatrix} \mathbf{s}_{k-1} \\ \mathbf{s}_k \end{bmatrix} \\ &= \mathbf{U}_k \mathbf{s}_{k-1} + \mathbf{L}_k \mathbf{s}_k, \end{aligned} \tag{3}$$

where $\mathbf{U}_k \mathbf{s}_{k-1}$ is the overlap from the previous frame to the current frame and $\mathbf{L}_k \mathbf{s}_{\mathbf{k}}$ is the contribution of the current frame to the residual. We can recover the speech signal $\mathbf{s}_k$ from the residual by

$$\begin{aligned} \mathbf{s}_k &= \mathbf{L}_k^{-1}(\mathbf{r}_k - \mathbf{U}_k \mathbf{s}_{k-1}) \\ &= \mathbf{L}_k^{-1}\mathbf{r}_k - \mathbf{L}_k^{-1}\mathbf{U}_k \mathbf{s}_{k-1}, \end{aligned} \tag{4}$$

where $\mathbf{L}_k^{-1}(\mathbf{U}_k \mathbf{s}_{k-1})$ is known as the zero input response (ZIR). It can be subtracted from the input signal, whereby we define the modified input signal

$$\hat{\mathbf{s}}_k = \mathbf{s}_k - \mathbf{U}_{k-1}\mathbf{r}_{k-1} = \mathbf{L}_k \mathbf{r}_k. \tag{5}$$

Using the modified input signal, the residual is

$$\mathbf{r}_k = \mathbf{L}_k^{-1} \hat{\mathbf{s}}_k. \tag{6}$$

Given the quantized residual $\mathring{\mathbf{r}}_k$, the SNR can be maximized, minimizing the mean square error (MSE)

$$\|\mathbf{s}_k - \mathring{\mathbf{s}}_k\|^2 = \left\|\mathbf{L}_k^{-1}\left(\mathbf{r}_k - \mathring{\mathbf{r}}_k - \mathbf{U}_k \mathbf{s}_{k-1} + \mathbf{U}_k \mathring{\mathbf{s}}_{k-1}\right)\right\|^2, \tag{7}$$

where $\mathring{\mathbf{s}}_k = \mathbf{L}_k^{-1}(\mathring{\mathbf{r}}_k - \mathbf{U}_k \mathring{\mathbf{s}}_{k-1})$ is the quantized output. Thus, the target residual when quantizing $\mathbf{s}_k$ is

$$\check{\mathbf{r}}_k = \mathbf{r}_k - \mathbf{U}_k(\mathbf{s}_{k-1} - \mathring{\mathbf{s}}_{k-1}), \tag{8}$$

which results in a feedback loop.

The ***basic FIR*** windowing scheme approximates the IIR filter operation, as performed in the case of the conventional IIR by an FIR filter. Thus, the impulse response of the IIR filter is used as coefficients for an FIR filter. As the impulse response of an IIR filter is by definition infinite, this results into an FIR filter of infinite length. Thus, big matrices have to be accepted in order to give identical results.

$$\mathbf{H}_k \mathbf{r}_k = \tilde{\mathbf{s}}_k = \begin{bmatrix} \tilde{\mathbf{s}}_{k,0} \\ \tilde{\mathbf{s}}_{k,1} \\ \vdots \end{bmatrix} = \begin{bmatrix} \mathbf{G}_k \mathbf{r}_k \\ \mathbf{T}_{k,1}\mathbf{r}_k \\ \mathbf{T}_{k,2}\mathbf{r}_k \\ \mathbf{T}_{k,3}\mathbf{r}_k \\ \vdots \end{bmatrix}. \tag{9}$$

The original signal can be reconstructed from the residual by

$$\mathbf{s}_k = \sum_{l=0}^{\infty} \tilde{\mathbf{s}}_{k-l,l} = \mathbf{L}_k \mathbf{r}_k + \sum_{l=1}^{\infty} \mathbf{T}_{k-l,l}\mathbf{r}_{k-l}. \tag{10}$$

By defining the ZIR as

$$\mathbf{q}_k = \sum_{l=1}^{\infty} \mathbf{T}_{k-l,l}\mathbf{r}_{k-l}, \tag{11}$$

we can define a modified input signal $\hat{\mathbf{s}}_k$ in the same manner as in the case of the conventional IIR

$$\hat{\mathbf{s}}_k = \mathbf{s}_k - \mathbf{q}_k = \mathbf{s}_k - \sum_{l=1}^{\infty} \mathbf{T}_{k-l,l}\mathbf{r}_{k-l} = \mathbf{G}_k\mathbf{r}_k. \tag{12}$$

Thus, the residual can be obtained from the modified input signal as

$$\mathbf{r}_k = \mathbf{G}_k^{-1}\hat{\mathbf{s}}_k. \tag{13}$$

To give a MMSE solution, we need to minimize

$$\|\mathbf{s}_k - \mathring{\mathbf{s}}_k\|^2 = \left\|\mathbf{G}_k\mathbf{r}_k + \mathbf{q}_k - \mathbf{G}_k\mathring{\mathbf{r}}_k - \mathring{\mathbf{q}}_k\right\|^2, \tag{14}$$

where $\mathbf{q}_k = \sum_{l=1}^{\infty} \mathbf{T}_{k-l,l}\mathbf{r}_{k-l}$. Thus, the target residual can be defined as

$$\check{\mathbf{r}}_k = \mathbf{r}_k + \mathbf{G}_k^{-1}(\mathbf{q}_k - \mathring{\mathbf{q}}_k). \tag{15}$$

In the case of the conventional IIR , the ZIR from the current frame, resulting from the current linear predictive filter is subtracted from the current frame. In contrast, for the case of the basic FIR the ZIR of the current frame with the current linear predictive filter is removed from the next frame. Analytically this is the only difference between the two methods.

The ***windowed FIR*** shortens the impulse response of the FIR filter by windowing it to an arbitrary length, to make it computationally feasible. To limit the overlap, that it only has influence on the next frame, the length was chosen to 2N. The output signal is obtained by filtering the residual with the $2\text{M} \times \text{M}$ convolution matrix $\mathbf{H}_k$

$$\mathbf{H}_k\mathbf{r}_k = \tilde{\mathbf{s}}_k = \begin{bmatrix} \tilde{\mathbf{s}}_{k,0} \\ \tilde{\mathbf{s}}_{k,1} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_k\mathbf{r}_k \\ \mathbf{T}_k\mathbf{r}_k \end{bmatrix}, \tag{16}$$

where $\mathbf{G}_k$ and $\mathbf{T}_k$ are upper- and lower-triangular Toeplitz matrices. The original signal can be obtained by overlap-add

$$\mathbf{s}_k = \tilde{\mathbf{s}}_{k,0} + \tilde{\mathbf{s}}_{k-1,1} = \mathbf{G}_k\mathbf{r}_k + \mathbf{T}_{k-1}\mathbf{r}_{k-1}. \tag{17}$$

By defining the ZIR as $\mathbf{q}_k = \mathbf{T}_{k-1}\mathbf{r}_{k-1}$ we can define $\hat{\mathbf{s}}_k$ as

$$\hat{\mathbf{s}}_k = \mathbf{s}_k - \mathbf{q}_k = \mathbf{s}_k - \mathbf{T}_{k-1}\mathbf{r}_{k-1} = \mathbf{G}_k\mathbf{r}_k, \tag{18}$$

whereby the residual is

$$\mathbf{r}_k = \mathbf{G}_k^{-1}\hat{\mathbf{s}}_k. \tag{19}$$

When the residual is quantized $\mathring{\mathbf{r}}_k = \mathbf{r}_k + \mathbf{n}_k$ the SNR can be maximized, minimizing the MSE

$$\begin{aligned}\|\mathbf{s}_k - \mathring{\mathbf{s}}_k\|^2 &= \|\mathbf{G}_k\mathbf{r}_k + \mathbf{T}_{k-1}\mathbf{r}_{k-1} - \mathbf{G}_k\mathring{\mathbf{r}}_k - \mathbf{T}_{k-1}\mathring{\mathbf{r}}_{k-1}\|^2 \\ &= \left\|\mathbf{G}_k\mathbf{r}_k + \mathbf{q}_k - \mathbf{G}_k\mathring{\mathbf{r}}_k - \mathring{\mathbf{q}}_k\right\|^2. \end{aligned} \tag{20}$$

Thus, the target residual can be defined as

$$\check{\mathbf{r}}_k = \mathbf{r}_k + \mathbf{G}_k^{-1}(\mathbf{q}_k - \mathring{\mathbf{q}}_k), \tag{21}$$

where $\mathbf{q}_k = \mathbf{T}_{k-1}\mathbf{r}_{k-1}$.

With ***windowed ZIR*** we refere to the approach that applies a window to the ZIR of the filter, in contrast to the *windowed FIR* , that windows the impulse response of the FIR filter. The windowed ZIR, can be implemented both, using FIR and IIR filters. For simplicity, we only give the formulation for the FIR filter version in the following. The re-synthesis is equivalent to the one presented in the basic FIR case.

$$\mathbf{H}_k\mathbf{r}_k = \tilde{\mathbf{s}}_k = \begin{bmatrix} \tilde{\mathbf{s}}_{k,0} \\ \tilde{\mathbf{s}}_{k,1} \\ \vdots \end{bmatrix} = \begin{bmatrix} \mathbf{G}_k\mathbf{r}_k \\ \mathbf{T}_{k,1}\mathbf{r}_k \\ \vdots \end{bmatrix}, \tag{22}$$

where $\mathbf{G}_k$ is a lower-triangular Toeplitz matrix and $\mathbf{T}_{k,h}$ are Toeplitz matrices. In the next step the ZIR is windowed

$$\widehat{\mathbf{s}}_k = \begin{bmatrix} \tilde{\mathbf{s}}_{k,0} \\ \mathbf{W}\tilde{\mathbf{s}}_{k,1} \end{bmatrix}, \tag{23}$$

where $\mathbf{W}$ is a diagonal matrix with diagonal entries ω_h, representing a continuous decreasing windowing function such that $1 > \omega_0 > \omega_1 > \cdots > \omega_{N-1} > 0$. For example, we can chose $\omega_k = \cos(\pi(k + 0.5)/2N)$.

The output signal is then defined by the overlap-add

$$\mathbf{s}_k = \tilde{\mathbf{s}}_{k,0} + \mathbf{W}\tilde{\mathbf{s}}_{k-1,1} = \mathbf{G}_k\mathbf{r}_k + \mathbf{W}\mathbf{T}_{k-1,1}\mathbf{r}_{k-1}. \tag{24}$$

We can then define the windowed ZIR as $\mathbf{q} = \mathbf{W}\mathbf{T}_{k-1,1}\mathbf{r}_{k-1}$. Thus we can define an input signal where the ZIR is removed

$$\hat{\mathbf{s}}_k = \mathbf{s}_k - \mathbf{q}_k = \mathbf{G}_k\mathbf{r}_k. \tag{25}$$

The residual can be calculated from $\hat{\mathbf{s}}_k$ as follows:

$$\mathbf{r}_k = \mathbf{G}_k^{-1}\hat{\mathbf{s}}_k = \mathbf{G}_k^{-1}\left(\mathbf{s}_k - \mathbf{W}\mathbf{T}_{k-1,1}\mathbf{r}_{k-1}\right). \tag{26}$$

In order to maximize SNR, with the definition
$\mathbf{q}_k = \mathbf{W}\mathbf{T}_{k-1,1}\mathbf{r}_{k-1}$, it follows that

$$\|\mathbf{s}_k - \mathring{\mathbf{s}}_k\|^2 = \left\|\mathbf{G}_k\mathbf{r}_k + \mathbf{q}_k - \mathbf{G}_k\mathring{\mathbf{r}}_k - \mathring{\mathbf{q}}_k\right\|^2. \tag{27}$$

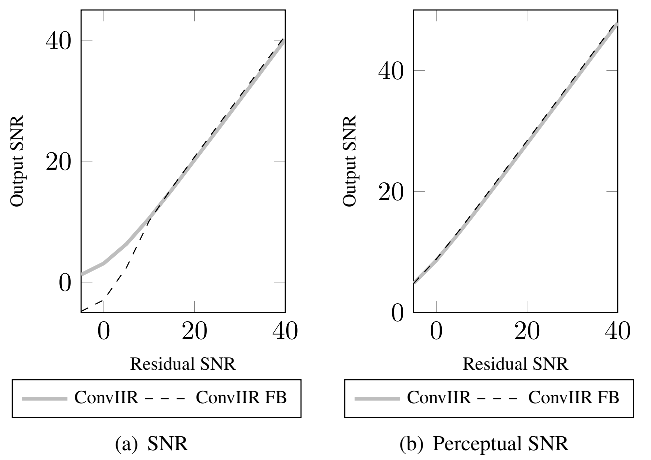
Therefore, the target residual $\check{\mathbf{r}}_k$ can be defined as

$$\check{\mathbf{r}}_k = \mathbf{r}_k + \mathbf{G}_k^{-1}(\mathbf{q}_k - \mathring{\mathbf{q}}_k), \tag{28}$$

resulting into a feedback loop, as also for the other approaches.

4. EXPERIMENTS

The purpose of this paper is to evaluate different approaches to filter-based windowing in the context of speech coding. For this evaluation, we chose to measure performance with



(a) SNR

(b) Perceptual SNR

Fig. 1. The objective measures for different quantization levels, presented for the conventional IIR with and without the feedback loop.

respect to quantization in the residual domain. The standard approach to quantization in speech codecs is to use the analysis-by-synthesis paradigm, where a large number of potential quantizations are evaluated iteratively to find the best codebook vector. However, since our objective is to study the inherent performance of windowing, we chose to omit the analysis-by-synthesis and used direct quantization of the residual instead. Specifically, to simulate quantization, we added a noise frame $\mathbf{n}_k$ to each residual frame $\mathbf{r}_k$ such that a constant SNR in the residual domain was achieved. Thus, the quantized residual is $\mathring{\mathbf{r}}_k = \mathbf{r}_k + \mathbf{n}_k$. To optimize output SNR, the residual has to be scaled, similarly as in algebraic code excited linear prediction (ACELP) codecs. Therefore, we scale the quantized residual with a coefficient γ, that is, $\hat{\mathbf{r}}_k = \gamma\mathring{\mathbf{r}}_k$ such that the output error $\|\mathbf{s}_k - \hat{\mathbf{s}}_k\|^2$ is minimized. The optimal γ can be readily found at the zero of the derivative, whereby the optimal γ for the IIR and FIR based approaches are, respectively

$$\gamma = \frac{\mathbf{r}_k^H \mathbf{L}_k^H \mathbf{L}_k \mathring{\mathbf{r}}_k}{\|\mathbf{L}_k \mathring{\mathbf{r}}_k\|^2} \quad \text{and} \quad \gamma = \frac{\mathbf{r}_k^H \mathbf{G}_k^H \mathbf{G}_k \mathring{\mathbf{r}}_k}{\|\mathbf{G}_k \mathring{\mathbf{r}}_k\|^2}. \tag{29}$$

The noise vector $\mathbf{n}_k$ was chosen as a random white noise vector for the objective tests, such that the SNR could be optimized. Since perceptual noise shaping is important for subjective tests, we filtered the random vector $\mathbf{n}_k$ with the same perceptual model as used in AMR-WB for all tests with a perceptual model as well as subjective tests [5].

The simulations were performed using a sampling rate of 12.8 kHz, a frame-length of 20 ms, and the length of the linear predictor was chosen to be 16, reflecting a typical wide-band speech coding scenario. The output SNRs were calculated averaging segmental SNR measure, based on a window length of 256. As audio samples for the evaluation, we used eight critical items of single channel speech and mixed material as used in the standardization of MPEG USAC [12].

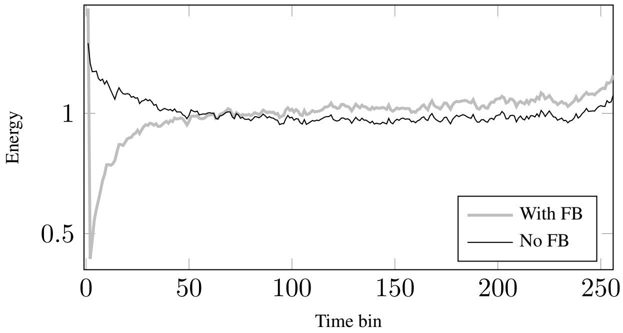


Fig. 2. The mean error energy per time bin of each frame, for the conventional IIR case, illustrating the time structure of the output error.

Figure 1(a) shows the result of the segmental SNR measured over different residual SNRs with and without the feedback loop for the case of the conventional IIR. It is evident that the performance of the windowing approach decreases at SNRs lower than 10 dB when no perceptual shaping is used together with the feedback loop. For the benefit of brevity we omitted the graphs of the other windowing schemes, as they showed no significant difference in any scenario. Therefore, the performance of all windowing scheme decreases for low SNRs when the feedback loop is used. Figure 1(b) shows the results of the perceptual SNR, for the case of the conventional IIR , applying perceptual shaping of the quantization noise. To determine if the output error has a temporal structure, we plotted the histogram of error energy of the output signal in Figure 2. It shows that, even though the SNRs are similar regardless whether the feedback loop is used, the temporal structure of the error varies greatly. Importantly, the approach applying the feedback loop has a large peak on the first sample, which produces a perceptually obvious temporal structure in the output signal. Also when applying perceptual shaping, all windowing approaches showed identical behaviour.

A MUSHRA test [13] was conducted with 12 listeners. In order to limit the listening test to a reasonable-length, the choice of items was limited to three due to the high amount of conditions under test. The presented items were processed according to the description of the objective tests. The test material was selected from the critical items list, used in standardization of MPEG USAC [12]: Item 1 and Item 3 are male_english and Item 2 is female_english. Quantization was simulated at an SNR of 0dB for Item 1 and 10dB for Items 2 and 3. The results of the listening test are shown in Figure 3. Item 1 clearly shows the trend that the feedback loop decreases perceptual quality at low SNR, whereas Items 2 and 3 show no significant effect of the feedback loop. The results show no other clear differences between the algorithms. For Item 1 a two way analysis of variance (ANOVA) ($\alpha = 0.05$) shows a significant main effect for the use of the feedback loop ($\mathrm{F}(1, 95) = 35.13, p <.001$). The average perceptual quality was higher without the feedback loop (M=45.25,SD=1.96),

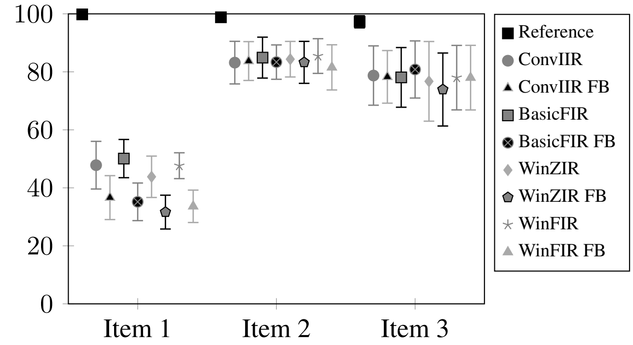


Fig. 3. The results of the MUSHRA listening test. Depicted are the means and the confidence intervals (95%).

than with the feedback loop (M=33.15,SD=1.77). However, the difference between the four presented algorithms was non-significant $(\mathrm{F}(3,95)=1, p=0.4)$. For Item 2 and 3 neither the usage of the feedback loop, nor the difference between the four presented algorithms were significant.

5. CONCLUSIONS

This paper presents three alternative windowing schemes, based on the filter-windowing technique were presented, differing in the transition between consecutive frames. In addition, each method can be combined with a feedback loop. The main result is that both objective and subjective results show that the feedback loop degrades quality when the quantization noise level is high. For low quantization noise levels the feedback loop improves output SNR by approximately 0.7dB. The presented results demonstrate that the performance of the alternative and the conventional windowing schemes are in typical scenarios effectively the same. The choice of windowing scheme can thus be based on other objectives, such as simplicity of design or ease of integration.

The conventional IIR, as used in speech coders is computationally inexpensive, but, due to the infinite ZIR tail, it suffers from a long error propagation. The same holds for the basic FIR approach, which moreover is computationally expensive as it requires the convolution with the filter response. Windowing the filters response to shorten its length yields the windowed FIR approach. This makes the approach computationally feasible, but has the disadvantage that it can result in an unstable filter, when it is inverted in the synthesis part. Moreover, changing the filter response will lower the potential prediction gain. The windowed ZIR approach offers a shorter error propagation while maintaining the prediction gain.

These results demonstrate that the conventional IIR and windowed ZIR methods offer best performance whereby the windowing scheme can be chosen to fit the overall system.

REFERENCES

[1] J. Benesty, M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, 2008.

[2] J. Allen, "Short-term spectral analysis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, pp. 235–238, 1977.

[3] W.B. Kleijn, D.J. Krasinski, and R.H. Ketchum, "Fast methods for the CELP speech coding algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 8, pp. 1330–1342, 1990.

[4] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. ICASSP*. IEEE, 1985, pp. 937–940.

[5] 3GPP, *TS 26.190, Adaptive Multi-Rate (AMR-WB) speech codec*, 2007.

[6] ITU-T G.718, "Frame error robust narrow-band and wideband embedded variable bit-rate coding of speech and audio from 8–32 kbit/s," *International Telecommunications Union, Geneva, Switzerland*, 2008.

[7] ISO/IEC 23003–3:2012, "MPEG-D (MPEG audio technologies), Part 3: Unified speech and audio coding," 2012.

[8] 3GPP, *TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12)*, 2014.

[9] III Barnwell, T.P., "Recursive windowing for generating autocorrelation coefficients for LPC analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 5, pp. 1062–1066, 1981.

[10] Juin-Hwey Chen, R.V. Cox, Y.-C. Lin, N. Jayant, and M.J. Melchner, "A low-delay CELP coder for the CCITT 16 kb/s speech coding standard," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 830–849, 1992.

[11] T. Bäckström, "Comparison of windowing in speech and audio coding," in *Proc. WASPAA*, New Paltz, USA, Oct. 2013.

[12] M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapierre, K. Kjörling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuiri, T. Chinen, T. Norimatsu, K. S. Chong, E. Oh, M. Kim, S. Quackenbush, and B. Grill, "The ISO/MPEG unified speech and audio coding standard – consistent high quality for all content types and at all bit rates," *Journal of the AES*, vol. 61, no. 12, pp. 956–977, 2013.

[13] ITU-R Recommendation BS.1534–2, "Method for the subjective assessment of intermediate quality level of audio systems (MUSHRA)," *International Telecommunications Union, Geneva, Switzerland*, June 2014.